

# Package ‘VariantToolsData’

December 26, 2024

**Type** Package

**Version** 1.30.0

**Maintainer** Michael Lawrence <michaf1a@gene.com>

**License** Artistic-2.0

**Title** Data for the VariantTools tutorial

**Author** Michael Lawrence

**Description** Data from the sequencing of a 50/50 mixture of HapMap trio samples NA12878 (CEU) and NA19240 (YRI), subset to the TP53 region.

**Suggests** VariantTools (>= 1.3.4), gmapR (>= 1.3.3), BiocStyle

**Imports** BiocGenerics, GenomicRanges

**Depends** R (>= 2.10), VariantAnnotation (>= 1.7.35)

**biocViews** ExperimentData, SequencingData, HapMap, SNPData

**git\_url** <https://git.bioconductor.org/packages/VariantToolsData>

**git\_branch** RELEASE\_3\_20

**git\_last\_commit** c6c4773

**git\_last\_commit\_date** 2024-10-29

**Repository** Bioconductor 3.20

**Date/Publication** 2024-12-26

## Contents

<b>geno</b> . . . . .	<b>2</b>
<b>repeats</b> . . . . .	<b>2</b>
<b>sumDepths</b> . . . . .	<b>3</b>
<b>tallies</b> . . . . .	<b>3</b>
<b>TP53Region</b> . . . . .	<b>4</b>
<b>Index</b>	<b>6</b>

---

geno	<i>Reference genotypes</i>
------	----------------------------

---

**Description**

Reference genotypes for NA12878 and NA19240, as called by the HapMap project, and the Broad GATK project.

**Usage**

```
data(geno)
```

**Format**

A VRanges with the genotypes for NA12878 and NA19240 from the HapMap pilot project and the Broad/GATK calling of NA12878. The genotypes, stored in the "freq." columns, are represented by the alt frequency, so 0/0.5/1 for hom-ref/het/hom-alt. The "expected.freq" column indicates the alt frequency expected in the 50/50 mixture.

**Source**

The HapMap Pilot and Broad GATK projects.

**Examples**

```
data(geno)
table(geno$expected.freq)
```

---

repeats	<i>Simple repeats</i>
---------	-----------------------

---

**Description**

Repeat regions from the RepeatMasker track of the hg19 UCSC genome browser database, subset to low complexity and simple repeats.

**Usage**

```
data(repeats)
```

**Format**

A GRanges object with the repeat ranges, including variables classifying the repeats by name, class, and family, and information about the alignment of the repeat consensus to the genome.

**Source**

The "rmsk" table in the UCSC table browser (hg19). Click the "Describe Schema" button for details.

**Examples**

```
data(repeats)
tab <- table(repeats$repFamily)
tab[tab > 0]
```

---

sumDepths

*Sum Replicate Depths*

---

**Description**

Finds the unique variants across every element of a list of VRanges, with depths computed by summing the depths over the elements. If a variant is not found in a sample, the depth is assumed to zero. That is a very dangerous assumption.

**Usage**

```
sumDepths(x)
```

**Arguments**

x                    A VRangesList, typically of replicates

**Value**

A VRanges

**Author(s)**

Michael Lawrence

**Examples**

```
data(tallies)
sumDepths(tallies)
```

---

tallies

*Tally VRanges*

---

**Description**

Nucleotide tallies computed over the TP53 region (+/- 1Mb) for the 50/50 NA12878/NA19240 mixture, separately for each replicate. Each replicate corresponds to a separate biochemical mixing.

**Usage**

```
data(tallies)
```

**Format**

A VRangesList, each VRanges of which corresponds to one of the three biochemical replicates.

**Source**

Computed from the alignments of the FASTQ files found in the 'inst/extdata' directory. Repeat regions (see [repeats](#)) were excluded. For example, for one replicate,

```
library(gmapR)
extdata.dir <- system.file("extdata",
                           package="VariantToolsData")
bams <- BamFileList(tools::list_files_with_exts(extdata.dir, "bam"))
data(repeats, package = "VariantToolsData")
param <- TallyVariantsParam(TP53Genome(), mask = repeats,
                            read_pos=TRUE, read_length=75L)
tallies <- split(tallyVariants(bams, param), ~ sampleNames)
```

This assumes that the BAM files have been generated for the current version of the TP53 genome:

```
param <- GsnapParam(TP53Genome(), unique_only = TRUE,
                   molecule = "DNA")
first.fastq <- dir(extdata.dir, "first.fastq",
                  full.names=TRUE)
last.fastq <- dir(extdata.dir, "last.fastq",
                 full.names=TRUE)
output <- gsnap(first.fastq, last.fastq, param)
bams <- as(output, "BamFileList")
```

**References**

Lawrence, M., Huntley, M. A., Stawiski, E., Owen, A., Wu, T. D., Goldstein, L. D., Cao, Y., Degenhardt, J., Young, J., Guillory, J., Heldens, S., Jackson, M., Seshagiri S., and Gentleman, R. (2015). Genomic variant calling: Flexible tools and a diagnostic data set. bioRxiv.

**Examples**

```
data(tallies)
VariantTools::callVariants(tallies[[1L]])
```

---

TP53Region

*Range around TP53*

---

**Description**

Returns a GRanges object consisting of the TP53 coordinates in hg19. All coordinates in these data are relative to that region.

**Usage**

```
TP53Region()
```

**Value**

A GRanges of the extents of the TP53 gene in hg19.

**Author(s)**

Michael Lawrence

**Examples**

TP53Region()

# Index

## \* datasets

  geno, [2](#)

  repeats, [2](#)

  tallies, [3](#)

geno, [2](#)

repeats, [2, 4](#)

sumDepths, [3](#)

tallies, [3](#)

TP53Region, [4](#)